

Sequences

Francis Ouellette
Director, CMMT Bioinformatics Core Facility
francis@cmmt.ubc.ca

Current Topics in Genome Analysis
Tuesday March 23, 1999

Outline

- What is Bioinformatics
- What is the primary sequence data
- What is GenBank
- Tools for submissions of sequences
- Updates: finishing the job

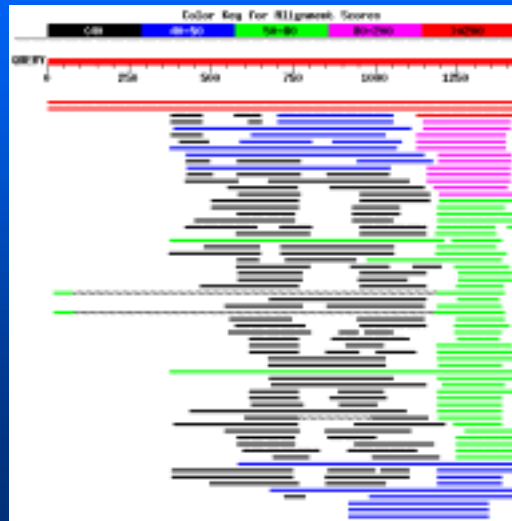
Computational Biology

- “New” Field of Science where mathematics, computer science and biology combine together to study and interpret genomic and proteomic information.
- CB will provide the tools for fully taking advantage of the HGP (est. 2003) as well as all of the other genome projects.
- CB will position its users at the head of the pack in any race for drug target discovery as well as improving healthcare worldwide.

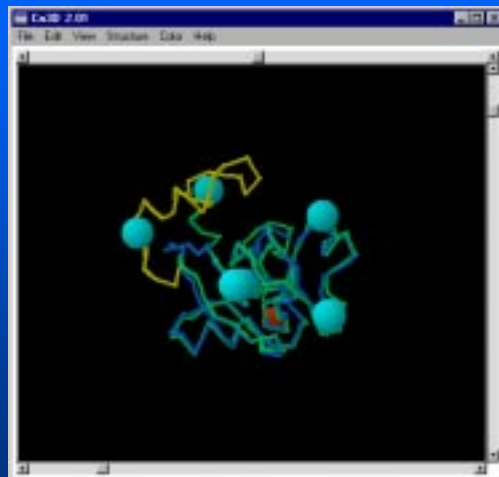
Bioinformatics is
about bringing biological
themes together with
the help of computer tools

BLAST Result

- Basic
- Local
- Alignment
- Search
- Tool



VAST result

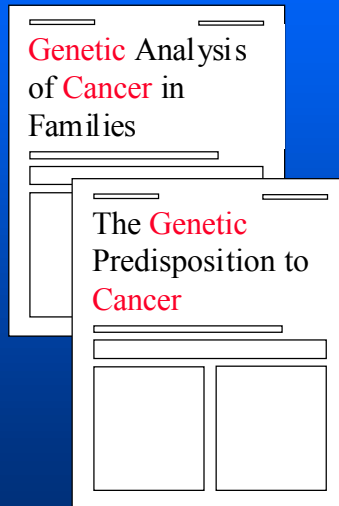


- Vector
- Alignment
- Search
- Tool

Ferredoxin

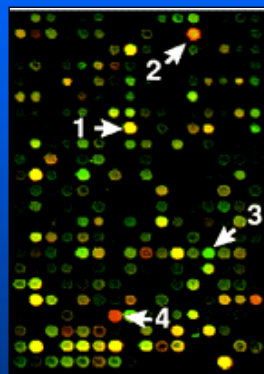
- *Halobacterium marismortui*
- *Chlorella fusca*

PubMed Text Neighboring



- Common terms could indicate similar subject matter
- Statistical method
- Weights based on term frequencies within document and within the database as a whole
- Some terms are better than others

Micro-array Analysis:



- mRNAs are more abundant in the serum-treated fibroblasts
- mRNAs are more abundant in the serum-deprived fibroblasts
- mRNAs does not vary substantially between the two samples

Iyer *et al.*, (1999) Science **283**: 83 - 87

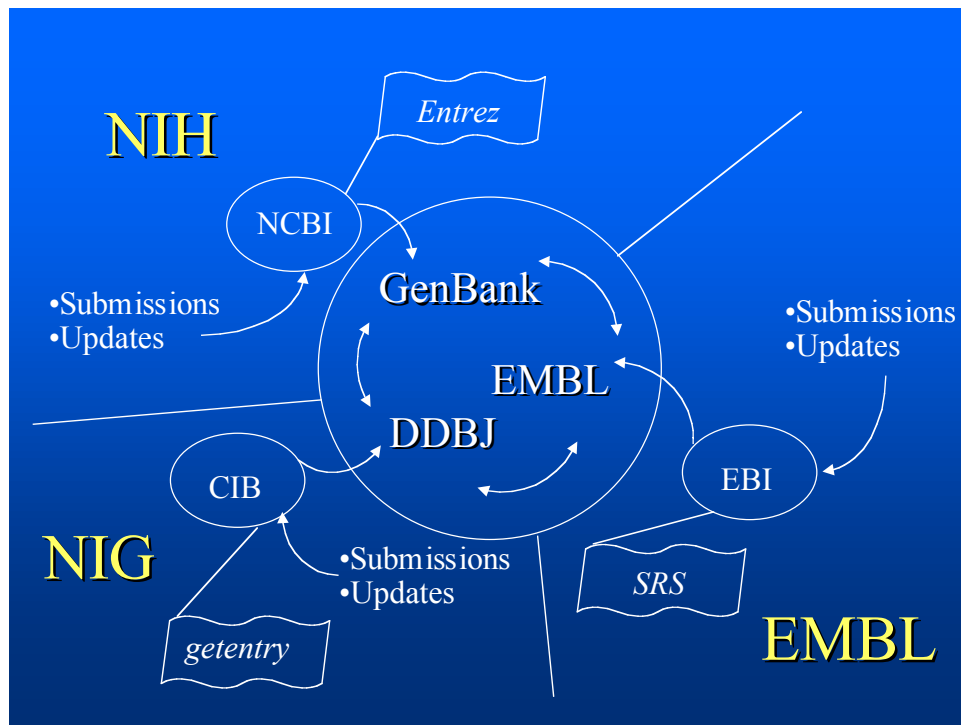
Primary Data

- DNA/RNA and protein sequences are the primary data for computational biology.
- Understanding the various types sequences present in GenBank is key to any interpretation in computational biology.
- Also understand that, as careful as NCBI and others are, errors do creep in, and always keep that critical eye open.

What is GenBank?

- GenBank is the NIH genetic sequence database of all publicly available DNA and derived protein sequences, with annotations describing the biological information these records contain.

<http://www.ncbi.nlm.nih.gov/Web/Genbank/index.html>
Benson *et al.*, 1999, *Nucleic Acids Res.* **27**:12-17



GenBank - Release 110 - Dec 1998

> 50,000 “species” or “terminal nodes”
 3,043,729 entries or GBFF
 2,162,067,871 nucleotides

- Full release of GenBank every 2 months.
- Incremental and cumulative releases: daily.
- GenBank is only available from the Internet.

GenBank Release 110 top 10

Records	Bases	Species
1,676,931	1,084,135,266	<i>Homo sapiens</i>
416,959	200,554,172	<i>Mus musculus</i>
76,740	147,625,327	<i>Caenorhabditis elegans</i>
71,363	82,625,620	<i>Arabidopsis thaliana</i>
67,852	72,625,772	<i>Drosophila melanogaster</i>
46,235	28,943,917	<i>Rattus norvegicus</i>
10,595	28,693,027	<i>Saccharomyces cerevisiae</i>
56,789	28,385,949	<i>Oryza sativa</i>
52,126	22,002,992	<i>Rattus sp.</i>
4,972	18,095,876	<i>Escherichia coli</i>
32,195	16,636,036	<i>Fugu rubripes</i>

Organismal Divisions

		Used in which database?
BCT	Bacterial	DDBJ - GenBank
FUN	Fungal	EMBL
HUM	Homo sapiens	DDBJ - EMBL
INV	Invertebrate	all
MAM	Other mammalian	all
ORG	Organelle	EMBL
PHG	Phage	all
PLN	Plant	all
PRI	Primate	all
PRO	Prokaryotic	EMBL
ROD	Rodent	all
RNA	Structural RNA	all
SYN	Synthetic and chimeric	all
VRL	Viral	all
VRT	Other vertebrate	all


Guiding Principals

- GenBank is a nucleotide-centric view of the information space.
- GenBank is a repository of all publicly available sequences.
- In GenBank, records are grouped for various reasons: understand this is key.
- Data in GenBank is only as good as what you put in: applying this is quite important.

GBFF and ASN.1

- GenBank data is maintained at the NCBI as ASN.1
- ASN.1 is a language that is used by computers to store, maintain, validate and show sequence information.
- the GenBank Flat File (GBFF) is one of these views (report) but has taken a life of its own.

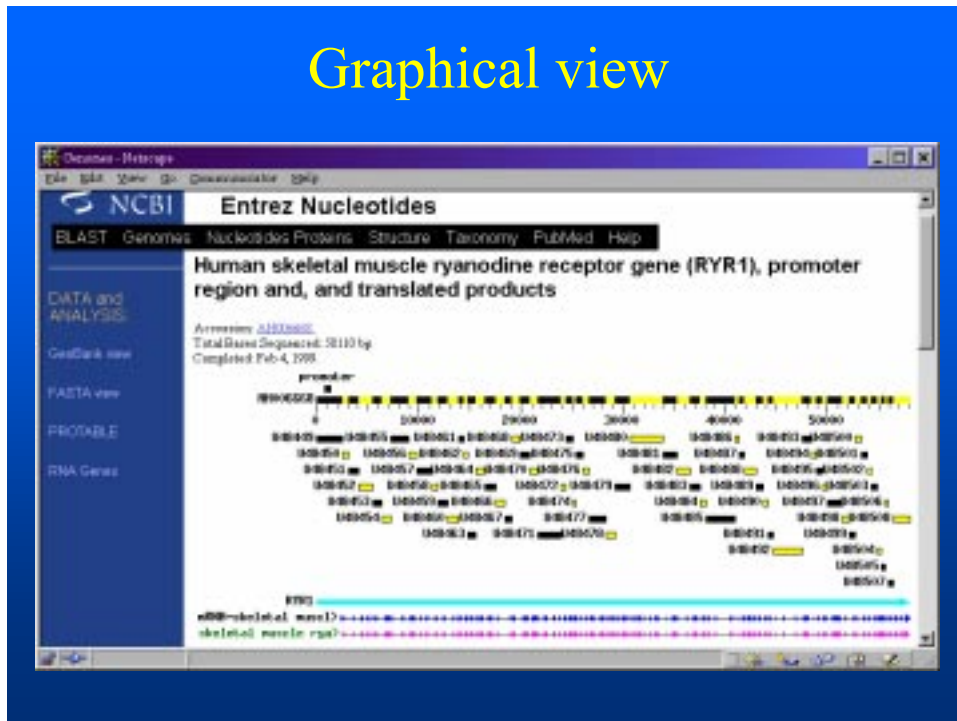
Abstract Syntax Notation (ASN.1)



```
seq-entry ::= set {
  class xso-prot ,
  -descr {
    create-date
    std {
      year 1999 ,
      month 7 ,
      day 17 ,
      title "Cucurbita maxima protein (PF16) mRNA, PF16-2 allele,
complete cds." ,
      source {
        -org {
          taxname "Cucurbita maxima" ,
          dbxref "vinter squash" ,
          -db {
            -db "taxdb" ,
            tag
            id 1993 : } ,
          -organism {
            name
            binomial {
              genus "Cucurbita" ,
              species "maxima" } ,
            mol {
              subtype cultivar ,
              subname "Big max" } : } ,

```

Graphical view



FASTA view

The screenshot shows the NCBI Entrez Nucleotide QUERY FASTA view for the Human skeletal muscle ryanodine receptor gene (RYR1). The window title is "Entrez - Nucleotide QUERY". The main title is "Nucleotide QUERY". The accession number is U11060. The total bases sequenced is 5810 by, and it was completed on Feb 4, 1995. The FASTA view displays the gene structure with exons represented by yellow boxes and introns by lines. The RYR1 gene is shown in green. The translated products are shown in blue and red. The window includes a menu bar with "BLAST", "Genomes", "Nucleotides", "Proteins", "Structure", "Taxonomy", "PubMed", and "Help". A sidebar on the left contains links for "DATA and ANALYSIS", "GenBank view", "FASTA view", "PROBLEMS", and "RNA Genes".

GenBank view



Sample GenBank Record

```

LOCUS       HSU40282     1786 bp     mRNA             PRI             28-NOV-1997
DEFINITION  Homo sapiens integrin-linked kinase (ILK) mRNA, complete cds.
ACCESSION   U40282
NID         g2648173
KEYWORDS    .
SOURCE      human.
  ORGANISM  Homo sapiens
            Eukaryotae; mitochondrial eukaryotes; Metazoa; Chordata;
            Vertebrata; Eutheria; Primates; Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1786)
AUTHORS     Hannigan,G.E., Leung-Hagsteijn,C., Fitz-Gibbon,L., Coppolino,M.G.,
            Radeva,G., Filmus,J., Bell,J.C. and Dedhar,S.
TITLE       Regulation of cell adhesion and anchorage-dependent growth by a new
            beta 1-integrin-linked protein kinase
JOURNAL     Nature 379 (6560), 91-96 (1996)
MEDLINE     96135142
REFERENCE   2  (bases 1 to 1786)
AUTHORS     Dedhar,S. and Hannigan,G.E.
TITLE       Direct Submission
JOURNAL     Submitted (07-NOV-1995) Shoukat Dedhar, Cancer Biology Research,
            Sunnybrook Health Science Centre and University of Toronto, 2075
            Bayview Avenue, North York, Ont. M4N 3M5, Canada
  
```

Sample GenBank Record: part 2

```

FEATURES             Location/Qualifiers
     source            1..1786
                        /organism="Homo sapiens"
                        /db_xref="taxon:9606"
                        /chromosome="11"
                        /map="11p15"
                        /cell_line="HeLa"
     gene              1..1786
                        /gene="ILK"
     CDS               157..1512
                        /gene="ILK"
                        /note="protein serine/threonine kinase"
                        /codon_start=1
                        /product="integrin-linked kinase"
                        /db_xref="PID:g2648174"
                        /translation="MDDIFTQCREGNAVAVRLWLDNTENDLNQGDDHGFSPLHWACRE
GRSAVVEMLIMRGARINVMNRGDDT PLHLAASHGHRDIVQKLLQYKAD INAVNEHGNV
... <deleted lines>
BASE COUNT          443 a    486 c    479 g    378 t
ORIGIN
      1 gaattcatct gtcgactgot accacgggag ttccccggag aaggatcctg cagcccgagt
... <deleted lines>
//

```

LOCUS, Accession, NID, gi and PID

LOCUS: Unique string of 10 letters and numbers in the database. Not maintained amongst databases, and is therefore a poor sequence identifier.

ACCESSION: A unique identifier to that record, citable entity; does not change when record is updated. A good record identifier, ideal for citation in publication.

NID: Nucleotide identifier: g, e or d prefix to gi number. Ideal identifier for specific version of a sequence.

Nucleotide gi: Geninfo identifier (gi), a unique integer which will change every time the sequence changes.

Accession.version: New system (expected late 1998) where the accession and version play the same function as the accession and gi number.

PID: Protein Identifier: g, e or d prefix to gi number. Can have one or two on one CDS.

Protein gi: Geninfo identifier (gi), a unique integer which will change every time the sequence changes.

protein_id: new identifier which will have the same structure and function as the nucleotide Accession and version numbers.

LOCUS, Accession, NID, gi and PID

LOCUS	HSU40282	1786 bp	mRNA	PRI	28-NOV-1997
DEFINITION	Homo sapiens integrin-linked kinase (ILK) mRNA, complete cds.				
ACCESSION	U40282				
NID	g2648173				

LOCUS: HSU40282
 ACCESSION: U40282
 NID: g2648173
 Nucleotide gi: 2648173
 VERSION: U40282.1 GI: 2648173
 PID: g2648174
 Protein gi: 2648174
 protein_id: AAA00000.1

CDS	157..1512
	/gene="ILK"
	/note="protein serine/threonine kinase"
	/codon_start=1
	/product="integrin-linked kinase"
	/db_xref="PID:g2648174"

NEW line types in GBFF

LOCUS	AAU36846	568 bp	DNA	PRI	26-OCT-1995
DEFINITION	Aotus azarai cytochrome c oxidase subunit II (COII) gene, mitochondrial gene encoding mitochondrial protein, partial cds.				
ACCESSION	U36846				
NID	g1040987				
VERSION	U36846.1 GI:1040987				
...					
CDS	<1..>568 /gene="COII" /codon_start=1 /product="cytochrome c oxidase subunit II" /protein_id = "AA12345.1" /db_xref="PID:g1040988" /db_xref="GI:1040988"				

GenBank - Release 110

<u>GB division</u>	<u>Nucleotides</u>
Organisms	933,008,136
EST	762,208,762
HTG	218,559,327
GSS	208,046,224
PAT	40,245,422
STS	22,535,866

GenBank Organismal divisions:

PRI - Primate	BCT - Bacterial
ROD - Rodent	RNA - Structural
MAM - Mammalian	VRL - Viral
VRT - Vertebrate	PHG - Phage
INV - Invertebrate	SYN - Synthetic
PLN - Plant	UNA - Unannotated

Functional Divisions

PAT - Patent

EST - Expressed Sequence Tags

STS - Sequence Tagged Sites

GSS - Genome Survey Sequences

HTG - High Throughput Genome

EST: Expressed sequence Tag

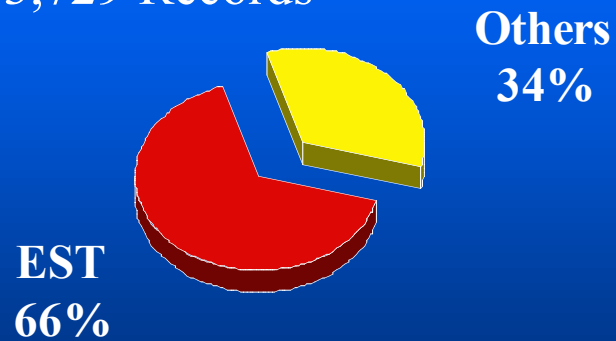
Expressed sequence Tags are short (300-500 bp) single reads from mRNA (cDNA) which are produced in large numbers.

They represent a snapshot of what is expressed in a given tissue, and developmental stage.

Also see: <http://www.ncbi.nlm.nih.gov/dbEST/>
<http://www.ncbi.nlm.nih.gov/UniGene/>

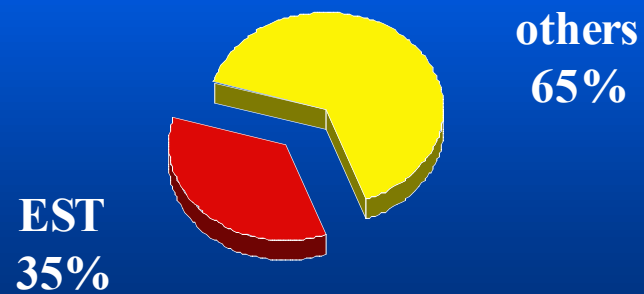
GenBank Release 110

3,043,729 Records

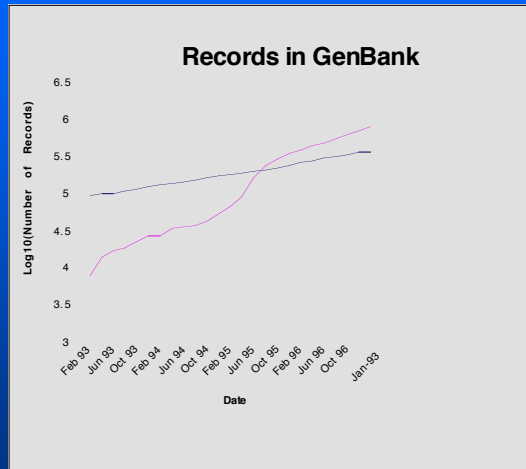


GenBank Release 110

2,162,067,871 Nucleotides



GenBank Growth



<ftp://ncbi.nlm.nih.gov/genbank/gbrel.txt>

```

LOCUS      AA675481      524 bp      mRNA      EST      28-NOV-1997
DEFINITION vr72d07.s1 Knowles Solter mouse 2 cell Mus musculus cDNA clone
            1134253 5' similar to TR:G992993 G992993 MYOSIN LIGHT CHAIN KINASE.
            ;.
ACCESSION  AA675481
NID        g2652718
KEYWORDS   EST.
SOURCE     house mouse.
  ORGANISM Mus musculus
            Eukaryotae; Metazoa; Chordata; Vertebrata; Mammalia; Eutheria;
            Rodentia; Sciurognathi; Muridae; Murinae; Mus.

...
COMMENT    Contact: Marra M/Mouse EST Project
            WashU-HHMI Mouse EST Project
            Washington University School of MedicineP
            4444 Forest Park Parkway, Box 8501, St. Louis, MO 63108
            Tel: 314 286 1800
            Fax: 314 286 1810
            Email: mouseest@watson.wustl.edu
            This clone is available royalty-free through LLNL ; contact the
            IMAGE Consortium (info@image.llnl.gov) for further information.
            MGI:615525
            Possible reversed clone: similarity on wrong strand
            High quality sequence stop: 469.
  
```

```

FEATURES             Location/Qualifiers
     source            1..524
                        /organism="Mus musculus"
                        /strain="B6D2 F1/J"
                        /note="Organ: embryo; Vector: pBluescribe (modified);
                        Site_1: MluI; Site_2: SalI; Cloned unidirectionally from
                        mRNA prepared from 13,500 2-cell stage embryos. Primer:
                        SalI(dT): 5'-CGGTCGACCGTCGACCGTTTTTTTTTTTTTTT-3'. cDNAs
                        were cloned into the MluI/SalI sites of a modified
                        pBluescribe vector using commercial linkers (NEB).
                        Average insert size: 1.2 kb."
                        /db_xref="taxon:10090"
                        /clone="1134253"
                        /clone_lib="Knowles Solter mouse 2 cell"
                        /tissue_type="embryo"
                        /dev_stage="2-cell"
                        /lab_host="DH10B"
BASE COUNT            168 a    111 c    115 g    130 t
ORIGIN
1   ctcagttgta gacagtgagc cagtcagatt tactgttaaa gtaacaggag aaccaagcc
61  ggaaattaca tgggtggttg aaggagaaat actgcaggat ggagaagact atcagtacat
121 cgaaagaggt gaaacttact gcctgtattt accggaaacc ttcccagaag atggaggaga
181 gtacatgtgt aaggcagtc acaataaagg ctcagcagcg agcacctgca ttcttaccat
241 tgaaatggat gactactagg cttccctctg tccttgggac tctctctctc gctgcatctc
301 tgtggagggg ccaaaaagga gaccagaggt gccactataa ctgacttaat ctttccccaa
361 atcttctct taagaacttc tcatgcata caggttcatt accatgctgt gcaaagtcaa
421 agcatagctg acagaaaagg gaaataaatg taccattctc gtcagaacta agacagaagc
481 ttcgtattta tagaactaag acttaacata tacagtttgc atga
//

```

STS

Sequenced Tagged Sites, are operationally unique sequence that identifies the combination of primer pairs used in a PCR assay that generate a mapping reagent which maps to a single position within the genome.

Also see: <http://www.ncbi.nlm.nih.gov/dbSTS/>
<http://www.ncbi.nlm.nih.gov/genemap/>

GSS

Genome Survey Sequences are similar in nature to the ESTs, except that its Sequences are genomic in origin, rather than cDNA (mRNA).

The GSS division contains:

- random "single pass read" genome survey Sequences.
- single pass reads from cosmid/BAC/YAC ends (these could be chromosome specific, but need not be)
- exon trapped genomic Sequences
- Alu PCR Sequences

Also see: <http://www.ncbi.nlm.nih.gov/dbGSS/>

HTG: High Throughput Genome

High Throughput Genome Sequences are unfinished genome sequencing efforts records. Unfinished records have gaps in the nucleotides sequence, low accuracy, and no annotations on the records.

Also see: <http://www.ncbi.nlm.nih.gov/HTGS/>

Ouellette and Boguski (1997) Genome Res. 7:952-955

HTGS in GenBank

phase 1 → ← → → ← HTG
 Acc = AC000003 gi = 1556454

phase 2 → → → HTG
 Acc = AC000003 gi = 2182283

phase 3 → PRI
 Acc = AC000003 gi = 2204282

40,000 to 120,000 bp

HTG: phase 1

```

LOCUS      HSAC000003 120000 bp      DNA           HTG           20-SEP-1996
DEFINITION *** UENCING IN PROGRESS *** Chromosome 17 genomic sequence; HTGS
            phase 1, 6 unordered pieces.
ACCESSION  AC000003
KEYWORDS   HTG; HTGS_PHASE1.
...
COMMENT    ***                                     ***
            *** WARNING: Phase 1 High Throughput Genome sequence ***
            ***                                     ***
            * This sequence is unfinished. It consists of 6 contigs for
            * which the order is not known; their order in this record is
            * arbitrary. In some cases, the exact lengths of the gaps
            * between the contigs are also unknown; these gaps are presented
            * as runs of N as a convenience only. When uencing is complete,
            * the sequence data presented in this record will be replaced
            * by a single finished sequence with the same accession number.
            *
            * 1      22526: contig of 22526 bp in length
            * 22527  23035: gap of unknown length
            * 23036  33919: contig of 10884 bp in length
            * 33920  34427: gap of unknown length
            * 34428  61877: contig of 27450 bp in length
            ...
            //
  
```

HTG: phase 3

```

LOCUS      AC000003      122228 bp      DNA      PRI      07-OCT-1997
DEFINITION Homo sapiens chromosome 17, clone 104H12, complete sequence.
ACCESSION  AC000003
NID        g2204282
KEYWORDS   HTG.
SOURCE     human.
...
COMMENT    The Staden databases, finishing information, and all
            chromatographic files used in the assembly of this clone are
            available from our anonymous ftp site.

            All repeats were identified using RepeatMasker: Smit, A.F.A. &
            Green, P. (1996-1997)
            http://ftp.genome.washington.edu/RM/RepeatMasker.html.
FEATURES             Location/Qualifiers
     source            1..122228
                        /organism="Homo sapiens"
                        /db_xref="taxon:9606"
                        /clone="104H12"
                        /clone_lib="Research Genetics/Cal Tech CITB978SK-B (plates
                        1-194)"
                        /chromosome="17"
     repeat_region     261..370
                        /rpt_family="MLT1B"

```

Guiding Principals

- In GenBank, records are grouped for various reasons, be it in organismal or functional divisions: understanding this is key to being able to fully exploit this database.

Why submit sequences to GenBank?

- No longer submit Sequences to Journal
- Journal scanning is no longer taking place
- Electronic format more useful and allows validations
- Sequences sent to DDBJ/EMBL/GenBank are exchanged daily
- Best way to exchange new data, and updates

What to submit?

- DNA sequence `atcttgctggggctgggtgaccaa`
- Where it comes from: The BioSource
`mouse mRNA from gene on chromosme 9`
- The citation, if one is present. The Cit-Sub
- What is relevent about this sequence? The annotated features.
- Other important Sequences: `encodes protein which has function Y in mouse development.`

Which Tool?

- BankIt: Web based tool which is simple, easy to use, great for simple submissions, but not ideal for complicated ones.
- Sequin: Client that you need to d/l to your computer, a little harder to learn, but has great documentation, and ideal for complicated, large, multiple submissions.

Sequin

- <http://www.ncbi.nlm.nih.gov/Sequin/>
- Sequence editor for new submissions or updates
- multi-platform (Mac/PC/Unix)
- built-in validation suite
- can do:
 - segmented sets
 - pop/phylo sets
 - large records
 - different views
 - specialized editors
 - complex or simple annotations
 - BLAST and *Entrez* client

Starting Sequin



Welcome to Sequin

Sequin

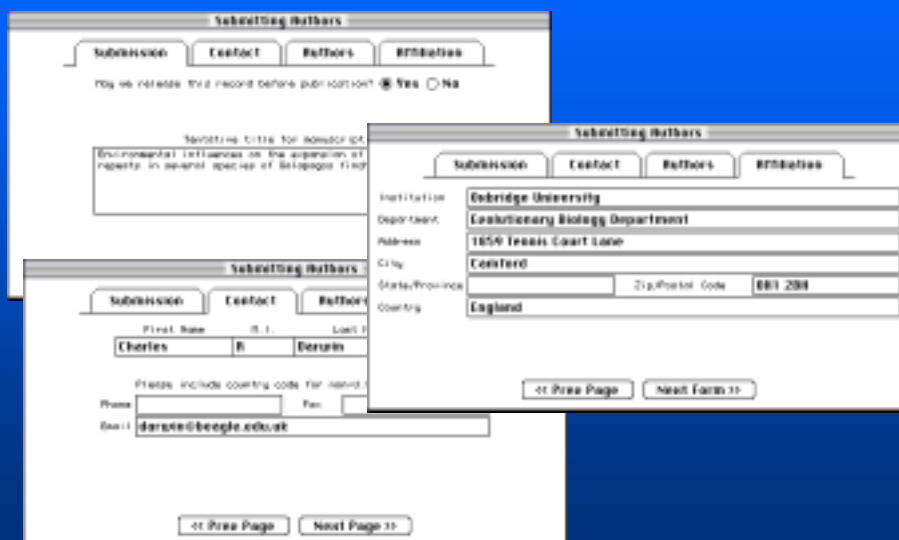
Sequin Application Version 2.22
Network Name: (Aug 13, 1997)

National Center for Biotechnology Information
National Library of Medicine
National Institutes of Health
13013-496-2475
info@ncbi.nlm.nih.gov

Database for submission: ☒ GenBank ☐ EMBL ☐ DDBJ

Start New Submission
Read Existing Record
Download from Entrez
Search Help
Exit Program

Author Information



Submitting Authors

Submission Contact Authors Attribution

May we release this record before publication? ☒ Yes ☐ No

Submitting title for reviewer (if Environmental releases on the expiration of patents in several species of biological kind)

Submitting Authors

Submission Contact Authors Attribution

Institution:
Department:
Address:
City:
State/Province:
Country:
Zip/Postal Code:

First Name: Initial: Last Name:

Please include country code for name:

Phone:
Fax:
Email:

Previous Page Next Page

DNA sequence Information



Sequence Format

☒ Single sequence ☐ Segmented sequence
 Submission type: ☐ Population study ☐ Phylogenetic study
☐ Mutation study ☐ Batch submission

Sequence data format: ☒ FRSTB ☐ FRSTB-GAP ☐ PIVLIP
☐ NDBS Interleaved ☐ NDBS Cologous



Organism and Sequences

Scientific Name:

Common Name:

Location of Sequence:

Use "Genomic" for a sequence encoded by a nuclear gene.

Genetic Code for Translation:

Adding Protein Sequences

FASTA file

```
>4E-I [gene=eIF4E] [prot=eukaryotic initiation factor 4E-I]
MQSDFHRMKNFANPKSMFKTSAPSTEQGRPEPPTSAAAPAEAKDVKPKEDPQETGEPAGN ...
>4E-II [gene=eIF4E] [prot=eukaryotic initiation factor 4E-II]
MVLLETEKTSAPSTEQGRPEPPTSAAAPAEAKDVKPKEDPQETGEPAGNTATTTAPAGDD ...
```



Organism and Sequences

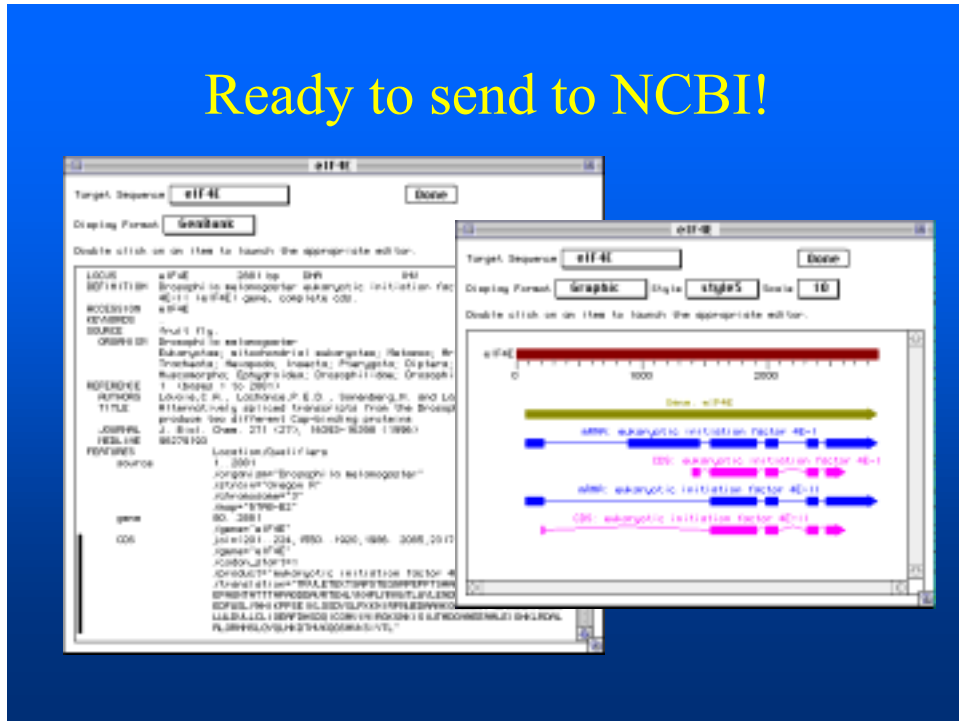
☐ Conceptual translation confirmed by peptide sequencing
☐ Incomplete at NDB end ☐ Incomplete at CDB end
☒ FRSTB def line starts with sequence ID
☒ Create initial index with CDB intervals

2 protein sequences, total length 587 amino acids

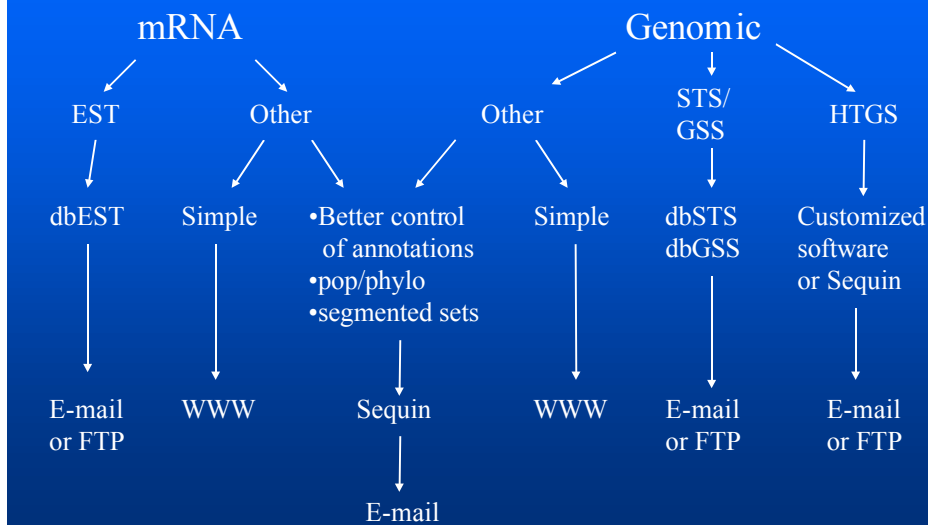
Choose Order from the Edit menu to clear these sequences

Sequence 1
 Length: 290 amino acids
 Sequence ID: 4E-I
 Gene: eIF4E
 Prot: eukaryotic initiation factor 4E-I
 No CDB detected

Ready to send to NCBI!



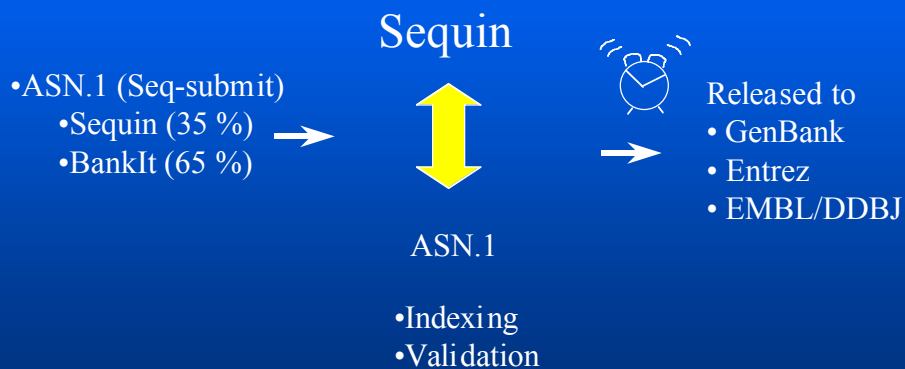
Which tool?



Where to Submit?

- Sequin files are e-mailed to:
gb-sub@ncbi.nlm.nih.gov
- BankIt: <http://www.ncbi.nlm.nih.gov/BankIt/>
- EST/GSS/STS
send e-mail to: batch-sub@ncbi.nlm.nih.gov
- HTGS
send query e-mail to: htgs@ncbi.nlm.nih.gov
- Not sure? e-mail to: info@ncbi.nlm.nih.gov

What NCBI does to your submission



What is validated?

■ Biological Validations

- gene name - division
- protein names - DEFINITION line

■ Computed validations

- Taxonomy - Splice sites
- CDS translate - Spelling
- Citation
- BLASTX/BLASTN (vector, duplications, updates)

Updates -- the 4 “W”

■ Who updates?

Submitters, Journals, “3rd party”

■ What to update?

Gene names, citations, new product, sequencing errors

■ Where?

update@ncbi.nlm.nih.gov

■ Why update?

BLAST Search Results - Ratings

File Edit View Go Command Window Help

Back Forward Reload Home Search Messages Find Security

Address: <http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/seqviewerblast>

What's Release

[BLAST - 1.9.0](#)

Distribution of 500 Blast Hits on the Query Sequence

P24942 CYTCB0808.8 gi|108177|gta||R1741.6 skiptape[...]-cphixbrom...B=61.3 E=2e-08

Color Key for Alignment Scores

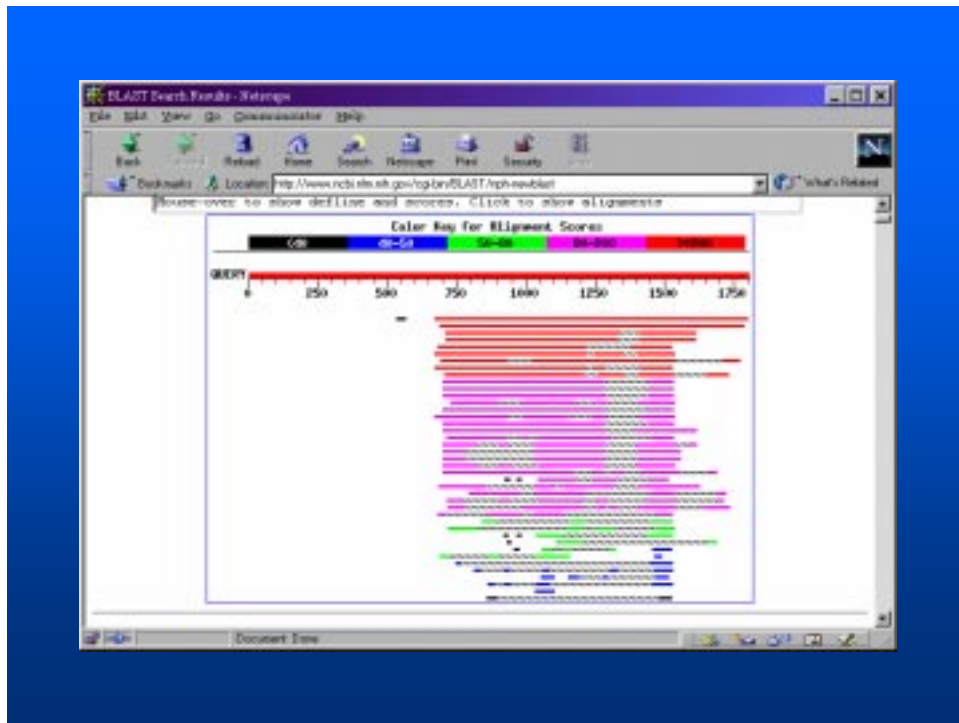
0-10 10-20 20-30 30-40 40-50 50-60

QUERY

0 250 500 750 1000 1250 1500 1750

<http://www.ncbi.nlm.nih.gov/cgi-bin/BLAST/seqviewerblast>

[illegible]



From francis@cmmt.ubc.ca Wed Mar 3 22:32:19 1999
To: ddbjupdt@ddbj.nig.ac.jp
Subject: D25291 mito

Dear colleagues,

it appears that DDBJ record D25291 is contaminated with mitochondrial Sequences from nucleotide 673 to 1803, as it is identical to mouse mitochondrial sequence (EMBL V00711) for more than 1100 nucleotides.

I would recommend deleting that segment of the record, or removing the record altogether, as it leads to unfortunate misinterpretation of the data when using GenBank or DDBJ. The protein sequence (which is erroneous, as it is all of mitochondrial origin) should definitely be removed as well.

Sequence Updated

```
LOCUS      MUSNGH      1803 bp      mRNA      ROD      29-AUG-1999
DEFINITION Mouse neuroblastoma and rat glioma hybridoma cell line NG108-1
            cell TA20 mRNA, complete cds.
ACCESSION  D25291
NID        g1850791
VERSION    D25291.1  GI:1850791
```

```
LOCUS      MUSNGH      619 bp      mRNA      ROD      12-MAR-1999
DEFINITION Mouse neuroblastoma and rat glioma hybridoma cell line NG108-1
            mRNA.
ACCESSION  D25291
NID        g9999999
VERSION    D25291.2  GI:9999999
```

Guiding Principals

- Data in GenBank is only as good as what you put in: applying and ensuring this (in an active, day too day fashion) will only make everybody's work that much more easier ...

Summary

- GenBank is a nucleotide-centric view of the information space, and is a report from the underlying ASN.1 data.
- In GenBank, records are grouped for various reasons: understand this is key to taking full advantage of this information.
- Sequin and BankIt can be used for updates and new submissions.
- Understanding the data elements in GenBank records is important, and allows you to take full control of the info.
- Data in GenBank is only as good as what you put in: applying this is crucial to the task at hand, dealing with the Everest of information before us.

Acknowledgments

GenBank Coordinator: Ilene Mizrachi

GenBank Annotation Staff:

Gabriella Ryan Adams, Medha Bhagwat, Lori Black, Larry Chlumsky, Karen Clark, **Irene Fang**, **Michael Fetchko**, Pamela Jacques, Anthony Jung, **Junga Kim**, Jaime Kolonay, Pierre Ledoux, Daniel Lyman, Baishali Maskeri, Jenny McDowell, Richard McVeigh, **Lillian Riddick**, Leigh Riley, Susan Schafer, **Jane Weisemann**, and Linda Yankie